Customer Segmentation Analysis Data-Driven Insights and Recommendations for AllLife Bank

Data Analysis and Visualization Elective Project.

By

Esmeralda C. Cabrera Ventura

February 16th, 2025.

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- Data Overview
- EDA and Data Preprocessing
- Model Building
- Appendix

Executive Summary

This report presents an analysis of AllLife Bank's credit card customers, and aims to identify customer segments based on financial behavior and interaction patterns. In this analysis, three clustering models were applied to segment customers. The first model was K-Means Clustering which created well-separated groups but was sensitive to outliers. Then, Gaussian Mixture Model (GMM) was applied, which allowed overlapping clusters, useful for nuanced segmentation although computationally expensive. Lastly, K-Medoids Clustering was applied to reduce the effect of outliers, resulting in more stable segmentation. By applying clustering techniques, three distinct customer groups were identified, enabling personalized marketing strategies and improved service delivery. The results showed that K-Medoids provided the most balanced segmentation, making it the most suitable model for business decisions. These findings suggest strategies for customer retention, credit-building opportunities, and digital banking adoption to drive business growth.

Business Problem Overview and Solution Approach

AllLife Bank wants to increase credit card market penetration while addressing customer concerns about support services. The primary objectives of this analysis include:

1. Segmenting customers based on credit limits, card ownership, and service interactions.
2. Identifying opportunities for targeted marketing and improved service efficiency.

To achieve these objectives, we applied clustering models such as K-Means, K-Medoids, and Gaussian Mixture Model (GMM) to categorize customers, offering actionable insights to optimize engagement strategies.

Data Overview

The dataset includes key attributes such as:

- Avg_Credit_Limit: The average credit limit assigned to customers.
- Total_Credit_Cards: The number of credit cards owned by each customer.
- Total_Visits_Bank: Frequency of in-person visits to the bank.
- Total_Visits_Online: Frequency of online banking interactions.
- Total_Calls_Made: Number of customer service interactions.

The dataset contains a total of 660 entries with 7 columns all of which are of numerical type. The are no missing values in the dataset used for this analysis. The information in the dataset was used to segment customers based on financial engagement and banking preferences. Other attributes such as SI_No and Customer Key were identifiers that were found to have a large number of unique values or to have duplicate values not useful for this report. More specifically, 5 duplicate customer keys and 11 duplicate rows were identified. Such duplicates were dropped and not considered in the analysis, bringing the number of relevant entries for our data analysis down to 644. Exact unique values for each attribute as shown in the table below.

Key Attributes Unique Value Table

Key Attribute	Unique Value Count
Sl_No	660
Customer Key	655
Avg_Credit_Limit	110
Total_Credit_Cards	10
Total_visits_bank	6
Total_visits_online	16
Total_calls_made	11

EDA and Data Preprocessing

Univariate Analysis

The table below contains the summary statistics of all the variables used in this analysis. Let's explore this information in more detail.

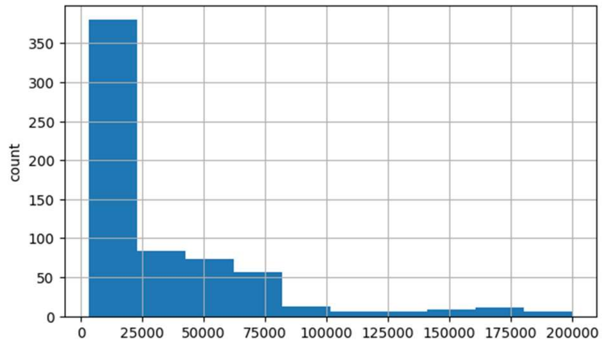
Summary Statistics

	count	mean	std	min	25%	50%	75%	max
Avg_Credit_Limit	644.0	34543.478261	37428.704286	3000.0	11000.0	18000.0	48000.00	200000.0
Total_Credit_Cards	644.0	4.694099	2.175338	1.0	3.0	5.0	6.00	10.0
Total_visits_bank	644.0	2.395963	1.626964	0.0	1.0	2.0	4.00	5.0
Total_visits_online	644.0	2.624224	2.957728	0.0	1.0	2.0	4.00	15.0
Total_calls_made	644.0	3.608696	2.880025	0.0	1.0	3.0	5.25	10.0

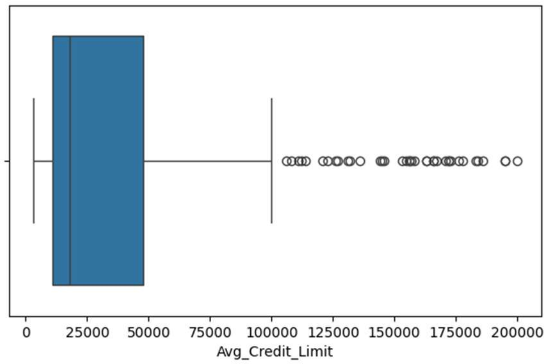
The summary statistics reveal key patterns in customer behavior. For instance, Avg_Credit_Limit varies widely, suggesting different customer segments such as high vs. low spenders. Overall, the average credit limit appears to be quite high, suggesting that the bank primarily serves customers with strong financial backgrounds. Total_Credit_Cards and interaction metrics such as bank visits, online visits, and calls show different banking habits with an average of 4.69 credit cards per customer.

High standard deviations and max values indicate there are some outliers, especially in credit limits and online visits and may need special handling. With regards to bank visits, online visits and calls made, the data suggests some customers frequently engage with the bank, while others rarely interact. From this, we deduce that customers can be segmented based on credit limits and banking interactions, and different service strategies can be applied to online users, frequent callers, and high-value customers. Let’s take a look at the resulting clusters from the data.

Average Credit Limit Histogram

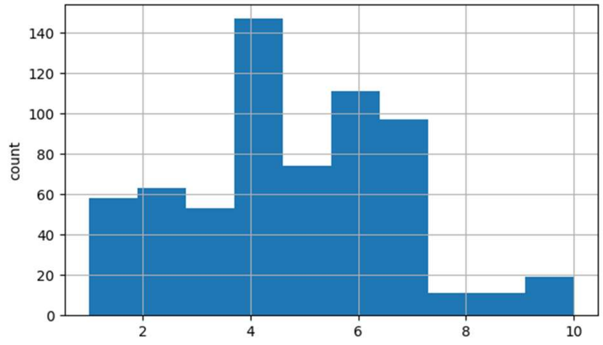


Average Credit Limit Boxplot

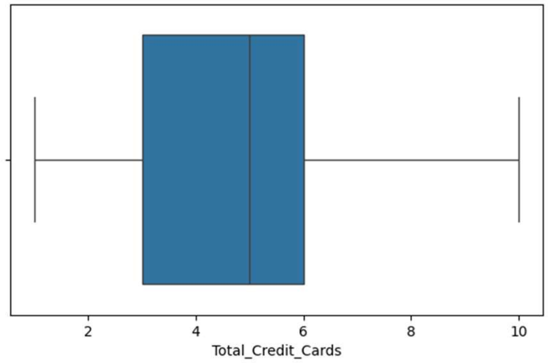


The histogram (left side) shows that most customers have low credit limits, while a small group of high-value customers has much higher limits, creating a right-skewed distribution. The boxplot (right side) confirms this by showing that the median is closer to the lower end, with many high-credit outliers. These outliers can either be analyzed separately as VIP customers or normalized to prevent them from skewing clustering results. To ensure accurate segmentation, clustering models should focus on the majority low-to-moderate credit group while making sure high-credit customers do not distort the results. Using scaling techniques will further improve clustering accuracy, leading to better customer targeting and personalized financial strategies.

Total Credit Cards Histogram



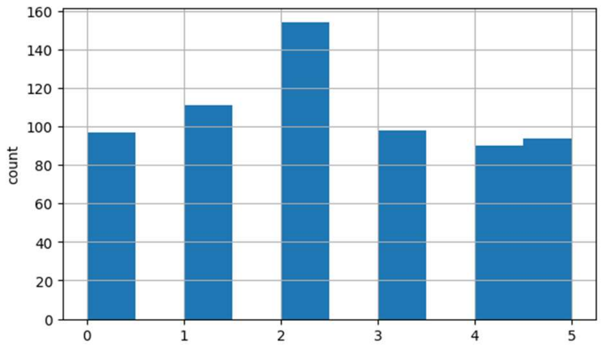
Total Credit Cards Boxplot



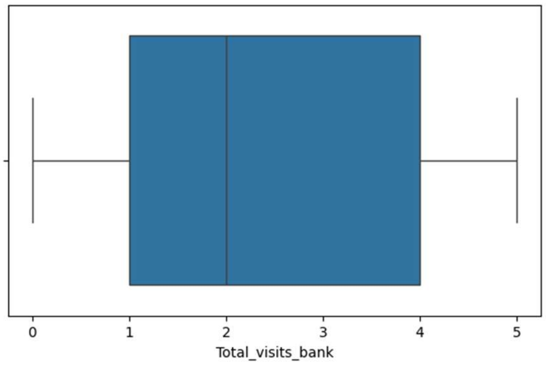
The histogram for Total Credit Cards shows two peaks, meaning most customers have either around 4 or 6 credit cards, with fewer having very low (1-2) or very high (9-10) card ownership. This suggests that customers naturally fall into two groups based on how many cards they own. The boxplot confirms this by showing that the median is near 5, meaning most customers have between 4 and 6 cards. Since the whiskers extend evenly and there are no major outliers, the data is well-balanced without extreme values.

For clustering, this means that credit card ownership is a reliable feature to separate customers into groups. Unlike credit limits, which required adjustments for extreme values, Total Credit Cards can be used directly in clustering models. When combined with the Average Credit Limit, it can help reveal different types of customers based on spending power and financial habits, making it useful for personalized banking strategies.

Total Bank Visits Histogram



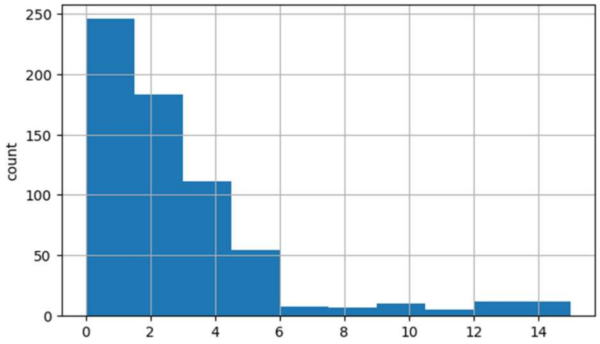
Total Bank Visits Boxplot



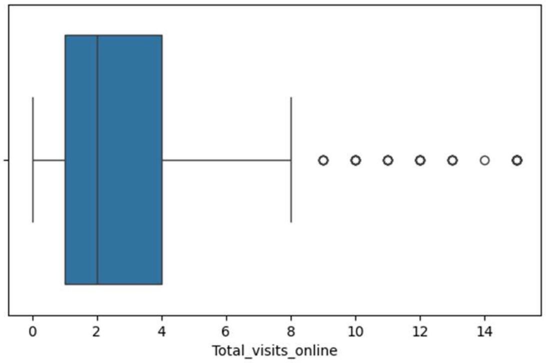
The histogram for Total Bank Visits shows that customer visits are evenly spread across all values (0 to 5), with the most common number of visits being 2, indicating that many customers engage in moderate in-person banking. A notable portion of customers never visit the bank (0 visits), suggesting that they prefer digital or phone services over in-person interactions. The boxplot confirms this trend, showing a median of around 2-3 visits, meaning that most customers visit the bank only a few times. Since the whiskers extend from 0 to 5 without any extreme values, the data is well-balanced with no significant outliers, reflecting different customer preferences for banking interactions.

For clustering, this feature can help classify customers based on how often they visit the bank. Non-visitors (0 visits) likely rely on online or phone banking, while moderate visitors (2-3 visits) use a mix of in-person and digital services. Frequent visitors (4-5 visits) may require more personalized, in-branch support. To improve segmentation, this feature should be combined with other banking interaction data (such as online visits and customer service calls) to accurately group customers based on their preferred service channels.

Total Online Visits Histogram



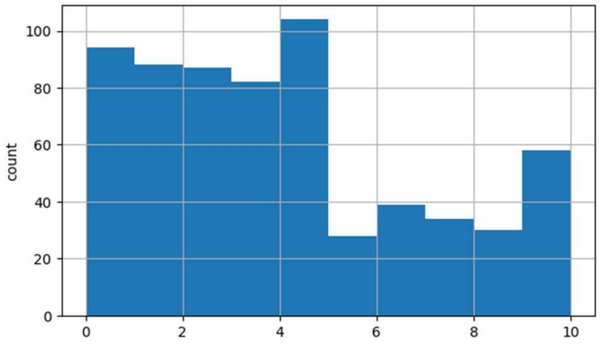
Total Online Visits Boxplot



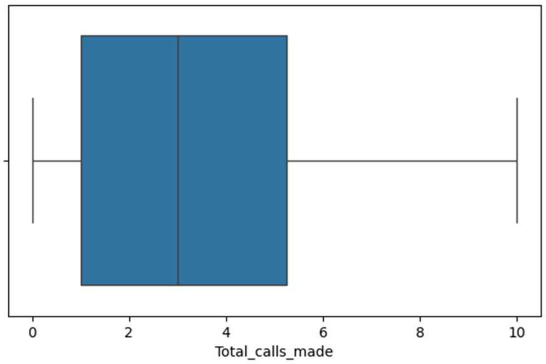
The histogram for Total Online Visits shows a right-skewed distribution, meaning most customers have low online engagement, with the highest frequency falling between 0-2 online visits. This suggests that a large portion of customers rarely or never use online banking services. However, a small percentage of customers visit online frequently (more than 8 times), indicating the presence of a distinct group of heavy online users who rely primarily on digital banking. The boxplot supports this finding, showing a median of around 3-4 visits, confirming that most customers engage in online banking only occasionally. The upper whisker extends to 6 visits, meaning that anything beyond this is considered an outlier. The presence of several outliers beyond 8 visits suggests that a small but significant group of customers is highly engaged in digital banking, using it far more than the average customer.

For clustering, customers can be segmented based on their level of digital engagement. Low online users (0-2 visits) likely prefer traditional banking methods such as in-person visits or phone support. Moderate users (3-6 visits) balance both online and offline services, while high online users (7+ visits, outliers) rely almost exclusively on digital banking. These tech-savvy customers could be targeted with digital-only promotions, mobile banking incentives, and self-service banking tools to enhance engagement and loyalty.

Total Calls Made Histogram



Total Calls Made Boxplot



The histogram for Total Calls Made shows a bimodal distribution, meaning customers tend to fall into two main groups: one that makes around 4 calls and another that makes around 10 calls. This suggests that some customers occasionally contact customer service, while others frequently rely on phone support. There is a noticeable dip between 6 and 8 calls, indicating that customers either call infrequently or very often, with fewer customers in the middle range. Additionally, a large portion of customers make very few calls (0-2 calls), suggesting they prefer alternative support channels such as online banking or in-person visits.

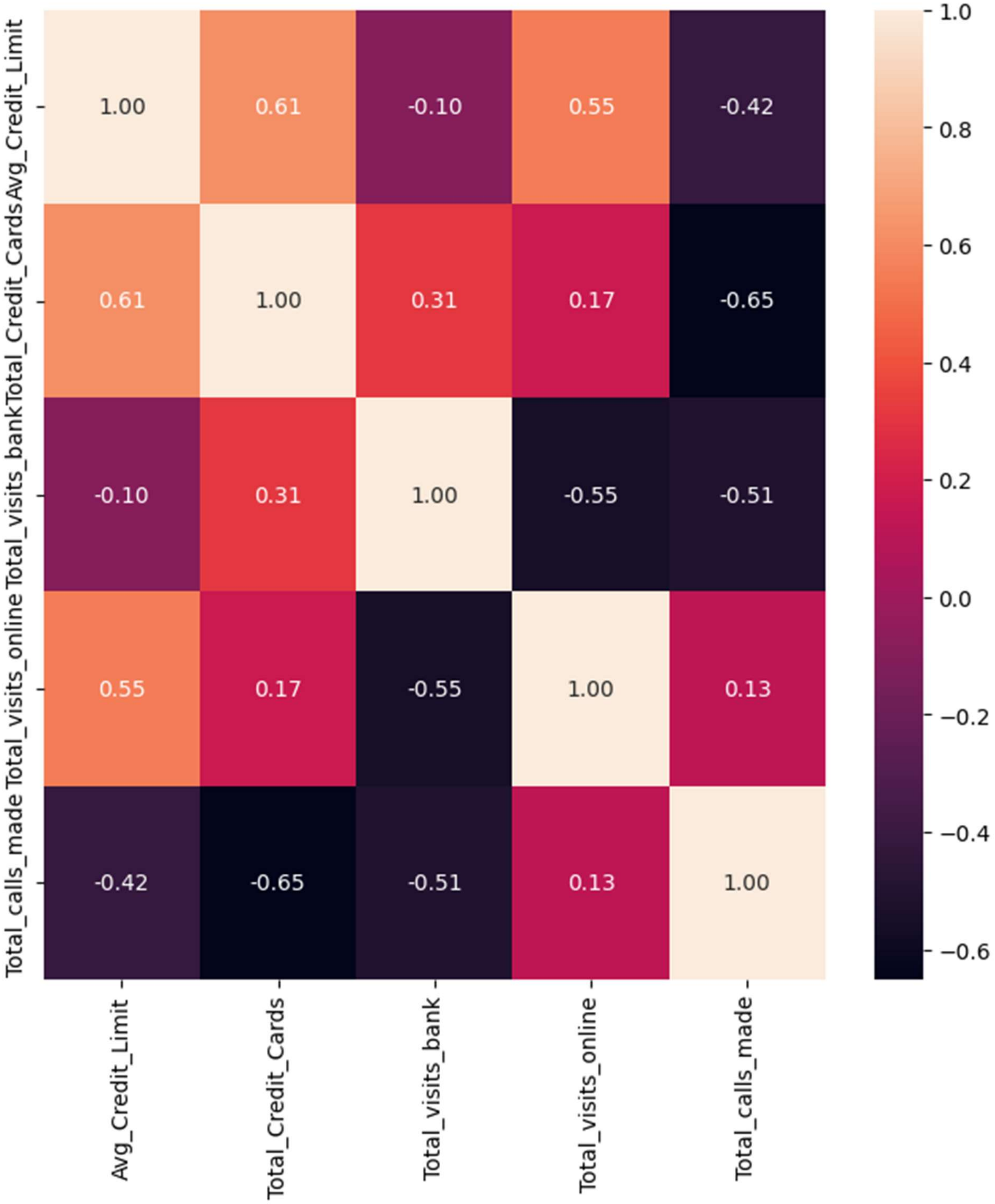
The boxplot confirms that the median number of calls is around 4-5, meaning that half of the customers make no more than these many calls. The whiskers extend from 0 to 10, showing that the data is evenly spread with no extreme outliers. The broad interquartile range (IQR) indicates wide variation in customer calling behavior, unlike online visits, which had clear outliers.

For clustering, this feature can help segment customers based on customer service preferences. Low callers (0-2 calls) likely prefer self-service options such as online banking or in-person visits. Moderate callers (3-5 calls) use a mix of phone support and other channels, while high callers (9-10 calls) rely heavily on phone support, potentially requiring better service optimization. Identifying these segments can help optimize call center resources, reduce wait times, and improve personalized service strategies for different customer needs.

Bivariate Analysis: Checking the Correlation

To examine the relationship between two variables at a time a correlation heatmap was used. In a correlation heatmap, each cell represents the correlation coefficient between a pair of variables, showing how strongly and in what direction (positive or negative) they are related. The correlation heatmap for the dataset is shown below. The heatmap provides valuable insights into how different variables in the dataset relate to each other. Correlation values range from -1 to 1, where a value close to 1 indicates a strong positive relationship (both variables increase together), a value close to -1 signifies a strong negative relationship (as one variable increases, the other decreases), and a value near 0 suggests no significant correlation. Let’s examine the key correlations in detail.

Variable Correlation Heatmap



Key correlations observed in the dataset highlight important customer behavior patterns. Avg_Credit_Limit and Total_Credit_Cards (+0.61) show a strong positive correlation, meaning that customers with more credit cards tend to have higher credit limits, which aligns with standard banking practices of increasing limits for customers with multiple credit lines. Similarly, Avg_Credit_Limit and Total_Visits_Online (+0.55) indicate that customers with higher credit limits use online banking more frequently, suggesting that high-credit customers prefer digital banking over physical visits. On the other hand, Total_Credit_Cards and Total_Calls_Made (-0.65) show a negative correlation, implying that customers with multiple credit cards require less phone support, possibly because they are more experienced with financial products.

Other significant relationships include Total_Visits_Bank and Total_Visits_Online (-0.55), confirming that customers who frequently visit the bank tend to use online banking less and vice versa, reinforcing the presence of two distinct customer segments: in-person vs. digital customers. Additionally, Total_Calls_Made and Avg_Credit_Limit (-0.42) suggest that lower-credit-limit customers make more calls, indicating that they might need more assistance with limit increases, balance inquiries, or account management.

These insights suggest that high-credit customers should be targeted with digital-first strategies, while frequent callers may need enhanced self-service tools. Since customers primarily interact through one channel (phone, online, or in-person), banks can optimize resources by tailoring services to each segment. Ultimately, these findings help improve customer engagement, banking service efficiency, and marketing strategies.

The correlation analysis revealed key patterns in customer behavior:

- Customers with higher credit limits tend to hold more credit cards.
- Online banking users have fewer in-person visits, showing a shift toward digital services.
- Frequent callers generally have lower credit limits, indicating a need for better financial support.

Additionally, outliers were detected in credit limits, requiring normalization before applying clustering models.

Model Building

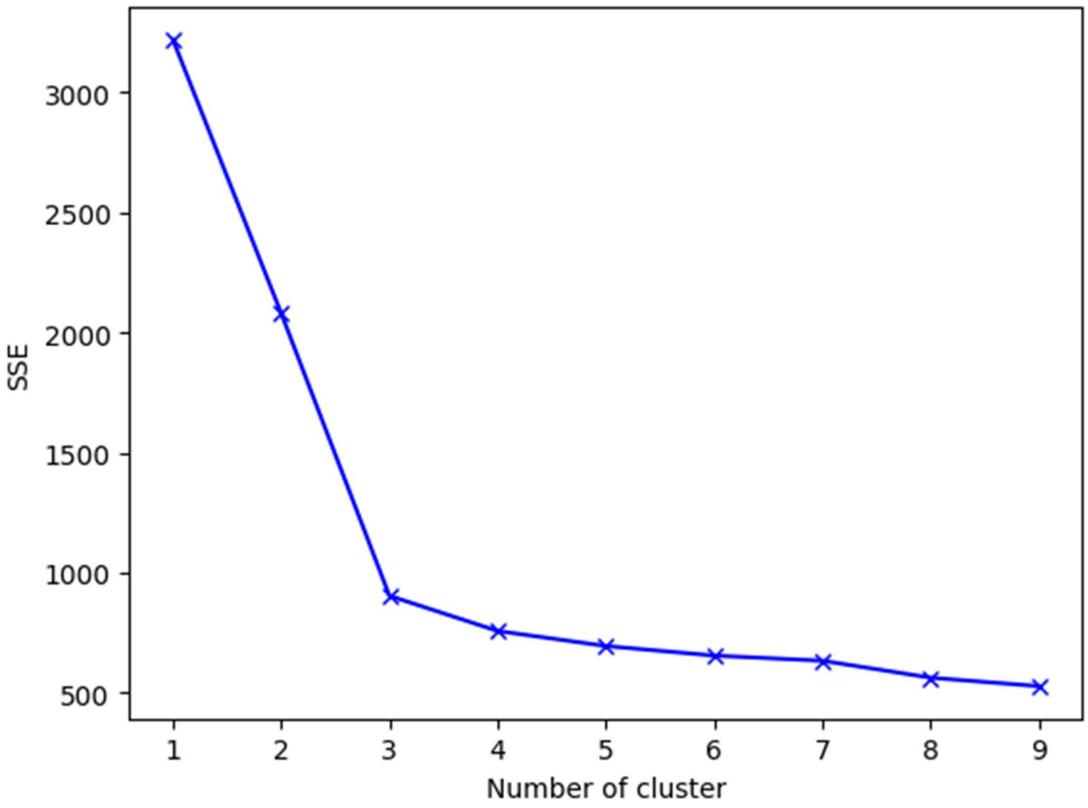
K-Means Model

Prior to modeling, it is necessary to scale the data. Scaling is applied before modeling to ensure that all features are treated fairly, especially in distance-based methods like K-Means, K-Medoids, and PCA. Without scaling, larger values (like credit limits) could dominate smaller ones (like bank visits), leading to biased results. It also improves PCA performance, makes outliers easier to detect, and helps models run more efficiently. Adjusting all features to a common scale (mean of 0, standard deviation of 1) ensures balanced comparisons, accurate clustering, and better model performance.

Once the data has been scaled, the next step is to apply PCA. PCA reduces redundancy, combines correlated variables, and simplifies high-dimensional data for better visualization. It also helps identify key features for clustering, detects customer behavior patterns, and confirms whether natural clusters exist before applying models like K-Means. By transforming the data into 2D or 3D, PCA ensures clustering is more accurate, interpretable, and focused on the most important attributes. Plotting PCA-transformed data, allows us to check if clusters are naturally present before applying modeling algorithms like K-Means as seen in the graph below.

The Elbow Method

K-Means Clustering Elbow



The Elbow Method is used to determine the optimal number of clusters (K) for K-Means clustering. It helps find the best number of clusters (K) for K-Means clustering by showing how the Sum of Squared Errors (SSE) decreases as more clusters are added. The X-axis represents the number of clusters tested, while the Y-axis shows how well the data fits into those clusters (lower SSE means better clustering). The graph forms an "elbow" shape, where SSE drops sharply until K = 3, then slows down, meaning adding more clusters doesn’t improve the results much.

The best choice for K in this case appears to be 3 because it significantly reduces SSE, creating clear and meaningful customer groups. Having fewer clusters (K = 1 or 2) means the data isn't well-separated.

. The K-Means algorithm yielded a total of 3 cluster profiles representing the group or segment to which each customer belongs based on shared characteristics. The cluster profile count (also referred to as cluster label) refers to the number of customers assigned to each cluster profile, group, or segment by the clustering algorithm for the dataset and is shown in the table below.

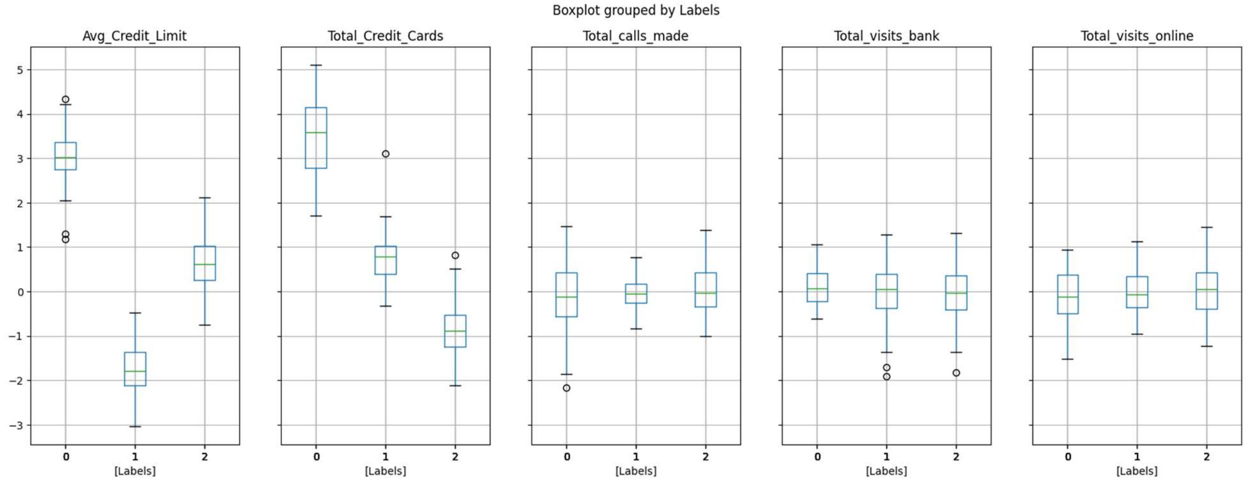
Cluster Profiles Table

Cluster Profile/Label	Count
2	374
1	221
0	49

The K-Means clustering results show that Cluster 2 is the largest with 374 customers, representing mid-range customers with moderate credit limits and engagement levels. This group presents an opportunity for upselling credit cards and personalized services. Cluster 1 (composed of 221 customers) is a significant segment of the customer base, possibly consisting of stable credit users who could benefit from retention programs and cross-selling financial products. Cluster 0, the smallest group (with 49 customers), is made up of high-value or outlier customers, such as those with very high credit limits or unique spending patterns. This segment should be targeted with exclusive perks, VIP programs, and customized financial planning.

To see how different customer clusters behave across key features, we can utilize summary statistics of the original data for each variable to plot a K-Means boxplot as shown below.

K-Means Label Based Boxplots



K-Means Clustering Profiles

The K-Means boxplots shown above illustrate how different customer clusters behave across key financial and banking interaction features. Cluster 0 represents high-value customers, characterized by the highest credit limits, multiple credit cards, and strong digital banking engagement, while making moderate use of customer support. These customers are likely VIPs or high-income individuals, and the bank should focus on exclusive rewards and premium services to retain them.

Cluster 1 consists of mid-tier customers, who have moderate credit limits and a balanced use of bank visits, calls, and online services. This segment represents the average customer base, making them ideal candidates for upselling financial products such as loans or savings plans. Cluster 2 includes low-value, high-support customers, with the lowest credit limits, the fewest credit cards, and the highest number of customer service calls, indicating a greater reliance on support services. These customers could benefit from financial education, self-service options, and credit-building opportunities.

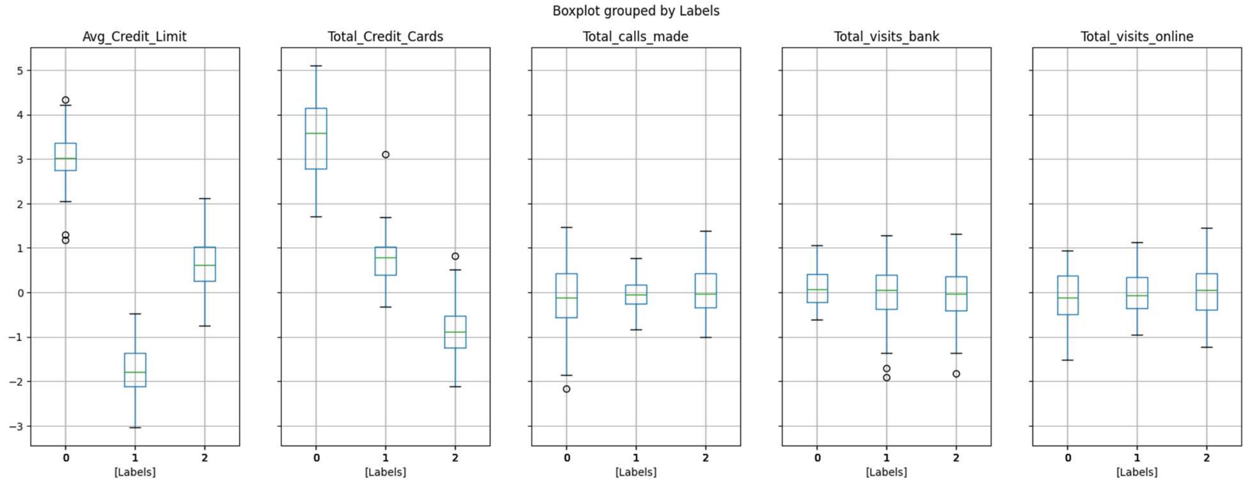
Gaussian Mixture Model (GMM)

Let's now create clusters using the Gaussian Mixture Model and see if we can extract more insights from our data. To this, the Gaussian Mixture algorithm was applied on the PCA components. However, when recalculating the summary statistics for each label in the original dataset with the GmmLabels, the results remained identical to that of the summary statistics used for K-Means clustering. Similarly, when plotting boxplots using the new GMM-based labels, the resulting visualizations are the same as those using K-Means labels. This suggests that the GMM clustering method did not significantly alter the overall distribution of key features as we can see in the table and boxplots below.

Summary Statistics

	group_0 Mean	**group_1 Mean**	**group_2 Mean**	**group_0 Median**	**group_1 Median**	**group_2 Median**
Avg_Credit_Limit	140102.040816	12239.819005	33893.048128	145000.0	12000.0	31500.0
Total_Credit_Cards	8.775510	2.411765	5.508021	9.0	2.0	6.0
Total_visits_bank	0.591837	0.945701	3.489305	1.0	1.0	3.0
Total_visits_online	10.979592	3.561086	0.975936	11.0	4.0	1.0
Total_calls_made	1.102041	6.891403	1.997326	1.0	7.0	2.0

K-Means Labels Based Boxplots



K-Medoids

The next step in the analysis is to apply the K-Medoids algorithm on the PCA components to see how different customer clusters behave across key features. The K-Medoids model produced the following cluster results:

Kmedoids Labels

KmedoLabels	Count
2	289
0	222
1	133

K-Medoids selects actual data points (medoids) as cluster centers, making it more stable and less affected by outliers. In this case, the K-Medoids algorithm divided customers into three groups based on their financial behavior. Cluster 2 (with 289 customers) is the largest group, includes all average customers with spending and banking habits, and making them a good target for upselling products and encouraging customer engagement. Cluster 0 (composed of 222 customers) represents a large but distinct group, representing financially stable customers with consistent banking habits, who might respond well to loyalty programs and exclusive offers.

Cluster 1 (with the last 133 customers) is the smallest group, composed of high-value customers or outliers with unique financial behaviors, such as premium credit card holders or big spenders. This group should be offered personalized financial products and VIP services to keep them engaged. By understanding these clusters, the bank can better tailor its marketing strategies and provide the right services to each group.

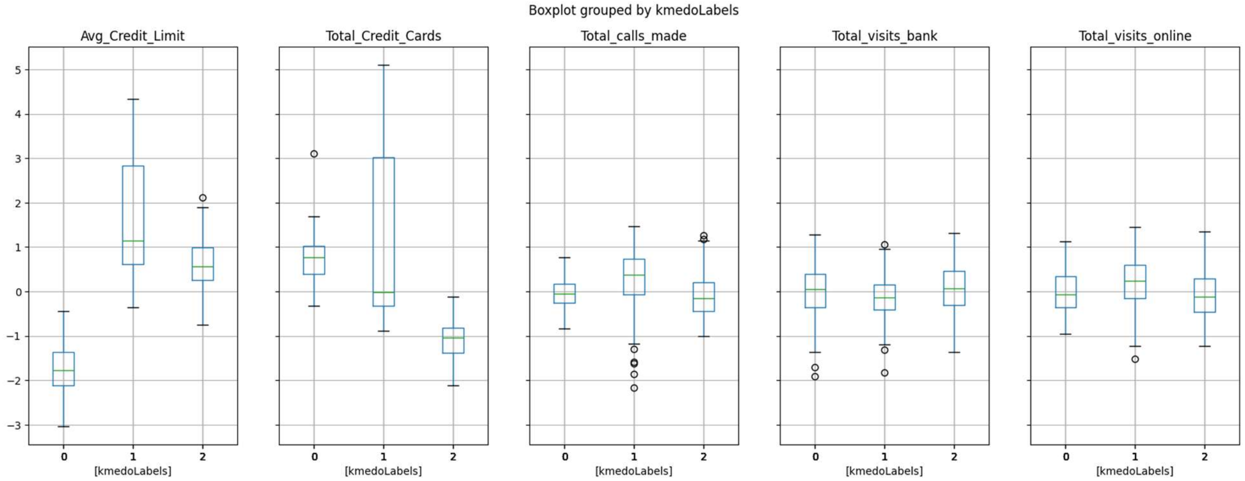
After applying K-Medoids and recalculating the summary statistics of the original data for each label, the following summary statistics are obtained:

K-Medoids Summary Statistics

	group_0 Mean	**group_1 Mean**	**group_2 Median**	**group_0 Median**	**group_1 Median**	**group_2 Median**
Avg_Credit_Limit	12216.216216	85052.631579	28449.826990	12000.0	68000.0	20000.0
Total_Credit_Cards	2.423423	7.030075	5.363322	2.0	7.0	5.0
Total_visits_bank	0.950450	1.691729	3.830450	1.0	2.0	4.0
Total_visits_online	3.554054	4.639098	0.982699	4.0	2.0	1.0
Total_calls_made	6.878378	1.969925	1.851211	7.0	2.0	2.0

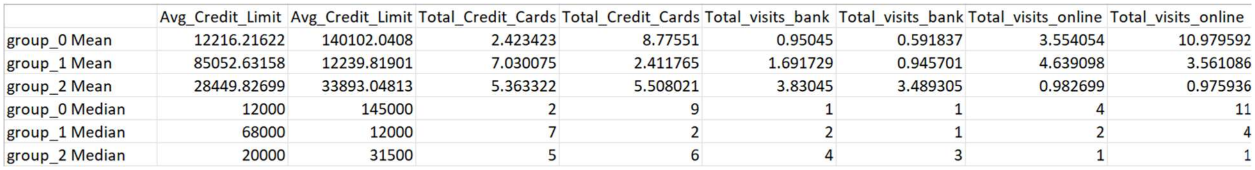
The summary statistics for each cluster from K-Medoids resulted in new values different from K-Means and GMM, providing key insights into customer segmentation. The values differ from K-Means and GMM due to the unique way K-Medoids assigns clusters, making it more robust to outliers and ensuring that the most representative data points (medoids) define each cluster.

K-Medoids Label Based Boxplots



The boxplots further highlight behavioral differences among the clusters. Cluster 1 stands out with the highest credit limits and financial engagement, making them ideal for premium banking services. Cluster 0 has the highest support needs and lowest credit limits, pointing to the need for better financial education and credit-building options. Cluster 2 visits branches most often, indicating a preference for personalized in-person interactions. While Cluster 0 is highly active online, Cluster 2 lags in digital banking usage, suggesting opportunities for digital transition incentives.

K-Means vs K-Medoids Clusters Comparison



	Avg_Credit_Limit	Avg_Credit_Limit	Total_Credit_Cards	Total_Credit_Cards	Total_visits_bank	Total_visits_bank	Total_visits_online	Total_visits_online
group_0 Mean	12216.21622	140102.0408	2.423423	8.77551	0.95045	0.591837	3.554054	10.979592
group_1 Mean	85052.63158	12239.81901	7.030075	2.411765	1.691729	0.945701	4.639098	3.561086
group_2 Mean	28449.82699	33893.04813	5.363322	5.508021	3.83045	3.489305	0.982699	0.975936
group_0 Median	12000	145000	2	9	1	1	4	11
group_1 Median	68000	12000	7	2	2	1	2	4
group_2 Median	20000	31500	5	6	4	3	1	1

When comparing K-Means and K-Medoids clusters, there were key differences in customer segmentation, banking interactions, and online behavior. K-Means produced more distinct clusters, clearly separating low, mid, and high-value customers, making it useful for marketing and personalized financial products. However, its sensitivity to outliers caused extreme values, like high-credit customers, to skew cluster assignments. In contrast, K-Medoids created more balanced clusters, reducing the influence of outliers and forming more stable customer segments that better reflect real-world behavior.

In terms of banking interactions, K-Means clearly distinguished digital-first vs. in-person customers, while K-Medoids captured a stronger digital segment, identifying customers with high online activity (10+ visits). Additionally, K-Medoids provided a more even distribution of customer service usage, while K-Means created sharper contrasts in call center reliance. This suggests that K-Means is more effective for marketing strategies, while K-Medoids is better suited for business decision-making and long-term engagement strategies.

Ultimately, K-Means is ideal for precise customer targeting and campaign-based strategies, while K-Medoids offers a more reliable segmentation for operational improvements and sustainable customer relationship management. The best choice depends on the bank’s objectives.

Conclusion

The analysis led to the identification of three key customer segments:

1. Low-credit, high-support, digital users requiring financial education and credit-building programs.
2. High-credit, low-support, balanced users requiring premium financial services and personalized incentives.
3. Mid-credit, high bank visits, traditional users that should be encouraged to adopt digital banking solutions.

The bank should align its strategies with the behavioral patterns identified in each cluster identified in this report. For example, digital-first customers should receive enhanced mobile banking features and personalized app-based services. High-support customers should be guided toward self-service tools and proactive customer education. Lastly but not the least, premium clients should be engaged with exclusive rewards programs and tailored financial offerings to ensure long-term loyalty. By implementing these targeted strategies, AllLife Bank can increase customer satisfaction, optimize marketing efforts, reduce operational costs, improve revenue streams, and operational efficiency.

Appendix

1. Data Preprocessing & Cleaning

Before applying clustering techniques, the dataset was cleaned and prepared for accurate segmentation.

- **Handling Duplicates:** The dataset initially contained 660 rows, but 16 were removed (5 due to being duplicates.
- **Missing Values:** No missing values were detected, so no imputation was required.
- **Scaling Data:** Clustering models rely on distance-based calculations, so StandardScaler() was applied to standardize numerical variables.
- **Outlier Detection:** Outliers in Avg_Credit_Limit and Total_Visits_Online were identified and considered for log transformation before clustering.

2. Model Selection & Justification

Three clustering algorithms were tested:

1- **K-Means Clustering** – Created distinct customer groups, useful for marketing but sensitive to outliers.

2- **Gaussian Mixture Model (GMM)** – Allowed clusters to overlap, making segmentation more flexible but slower to process.

3- **K-Medoids Clustering** – Selected actual customer data points as medoids, making clusters more stable and less affected by outliers.

3. Key Visualizations & Insights

- **Histogram & Boxplot Analysis**: Identified right-skewed credit limit distributions, showing most customers have low-to-moderate limits, with a small VIP segment.
- **Elbow Method Graph**: Determined K = 3 clusters as the optimal segmentation.
- **Correlation Heatmap**: Showed what variables that strong relationships
- **Cluster-Based Boxplots**: Compared customer spending, service usage, and banking interactions across clusters.

4. **Code Snippets for Key Analytical Steps after importing appropriate libraries**

Scaling Data:

```
scaler = StandardScaler()

data_scaled = scaler.fit_transform(data)
```

Elbow Method for Optimal Clusters:

```
sse = {}

for k in range(1, 10):

    kmeans = KMeans(n_clusters=k, max_iter=1000, random_state=1).fit(data_pca)

    sse[k] = kmeans.inertia_

plt.figure()

plt.plot(list(sse.keys()), list(sse.values()), 'bx-')

plt.xlabel("Number of Clusters")

plt.ylabel("SSE")

plt.show()
```

Applying K-Medoids Clustering:

```
from sklearn_extra.cluster import KMedoids

kmedoids = KMedoids(n_clusters=3, random_state=1).fit(data_pca)

data["Labels"] = kmedoids.labels_
```

Plotting Customer Segmentation Boxplots:

```
data.boxplot(by="Labels", layout=(1, 5), figsize=(20, 7))

plt.show()
```